CLAIMS

What is claimed is:

1. A system to provide finer grain control in optimizing multiple workloads across multiple servers, comprising:

a plurality of servers to be utilized by multiple workloads;

a plurality of virtual machines (VMs) at each of the plurality of servers, wherein the plurality of VMs at each of the plurality of servers each serve a different one of the multiple workloads; and

resource management logic to distribute server resources to each of the plurality of VMs according to predicted resource needs of each of the multiple workloads,

whereby, each of the multiple workloads are distributed across the plurality of servers, wherein fractions of each of the multiple workloads are handled by the plurality of VMs,

whereby, the fractions of each of the multiple workloads handled by each of the VMs can be dynamically adjusted to provide for optimization of the multiple workloads across the multiple servers.

2. The system of claim 1 wherein the distribution of server resources to each of the plurality of VMs further comprises distributing server resources to the plurality of VMs according to the current and predicted resource needs of each of the multiple servers.

3. The system of claim 2 wherein the server resources comprise percentage of CPU, percentage of network bandwidth, disk resources and memory resources.

4. The system of claim 1 wherein the finer grain control is achieved through recognizing when one of the plurality of servers is overloaded and shifting work to another of the plurality of servers which is not overloaded.

5. The system of claim 1 wherein the fractions of the multiple workloads being handled by the plurality of VMs can be dynamically adjusted in response to workload changes at the plurality of servers, wherein the dynamic adjustment provides for maintaining an optimum utilization level across the plurality of servers.

6. The system of claim 5 wherein the optimum utilization level can be configured automatically via server management software or manually by a user with administrative privileges.

7. The system of claim 1 wherein the workloads are each distributed over a subset of the plurality of VMs.

8. The system of claim 7 wherein each VM in the subset of the plurality of VMs exists at a separate one of the plurality of servers.

9. The system of claim 8 wherein the workload distribution comprises distributing the work according to resources available to each of the VMs within the subset.

10. The system of claim 1 further comprises at least one global resource allocator to monitor resource distribution between the plurality of VMs.

11. The system of claim 10 further comprises at least one load balancer to measure the current offered load.

12. The system of claim 11 wherein the global resource allocator determines how to distribute the resources between the plurality of VMs, according to the measurements received from the at least one load balancer.

13. The system of claim 12 wherein each of the plurality of servers includes a local resource control agent to receive and implement instructions from the global resource allocator describing how the resources are to be distributed between the VMs located at each of the plurality of servers.

14. A server optimization device, for providing finer grain control in a virtual machine (VM) based hosting architecture, comprising:

at least one load balancer component to identify resource requirements for multiple different workloads in the VM based hosting architecture;

a global resource allocator partitioning component to assign VMs from multiple server machines to a workload according to the identified resource requirements; and

the global resource allocator partitioning component to assign resources at each of the multiple server machines to the assigned VMs according to the identified resource requirements.

15. The server optimization device of claim 14 further comprises the global resource allocator partitioning component reassigning the VMs according changes in the identified resource requirements.

16. The server optimization device of claim 14 further comprises a plurality of resource allocator components at each of the multiple server machines, wherein the plurality of resource allocators are responsible for creating VMs and assigning VMs to workloads in response to instructions received from the global resource allocator partitioning component.

17. The server optimization device of claim 14 wherein the at least one load balancer continuously monitors the resource requirements for the multiple different workloads and provides changes to the resource requirements of each of the multiple different workloads to the global resource allocator partitioning component.

18. The server optimization device of claim 17 further comprises the global resource allocator partitioning component issuing instructions to the plurality of resource allocator

components at each of the multiple server machines, wherein the issued instructions provide for redistributing server resources to each of the VMs within each of the multiple server machines,

whereby, the redistribution of the server resources provides for optimizing workload across the multiple server to prevent the over-utilization or under-utilization of the multiple server machines.

19. The server optimization device of claim 14 wherein the VMs at each of the multiple server machines serve a different one of the multiple different workloads.

20. The server optimization device of claim 14 wherein the resources comprise percentage of CPU, percentage of network bandwidth, disk resources and memory resources.

21. The server optimization device of claim 18 further comprises the instructions being issued automatically via server management software in order to maintain a pre-defined level of optimization within the system.

22. The server optimization device of claim 14 wherein the multiple different workloads are distributed over a subset of the assigned VMs.

23. The server optimization device of claim 14 wherein the multiple different workloads are each assigned to a customer application utilizing the VM based hosting architecture.

24. A method for improving server utilization levels, comprising:

identifying a current offered load of each of a plurality of customer applications;

analyzing the identified current offered load and generating a prediction as to what resources will be needed by each of the plurality of customer applications;

identifying all VMs associated with each of the plurality of customer applications; and

allocating, according to the generated prediction, the resources needed by each of the plurality of customer applications to the VMs associated with each of the plurality of customer applications,

whereby, the offered load associated each of the customer applications is distributed over a plurality of the identified VMs, and the VMs over which each of the customer applications offered load is distributed, reside on separate machines.

25. A server optimization means, for providing finer grain control in a virtual machine (VM) based hosting architecture, comprising:

a means for identifying resource requirements for multiple different workloads in the VM based hosting architecture;

a means for assigning VMs from multiple server machines to a workload according to the identified resource requirements; and

a means for assigning resources at each of the multiple server machines to the assigned VMs according to the identified resource requirements.

26. The server optimization means of claim 25 further comprises a means for reassigning the VMs according changes in the identified resource requirements.

27. The server optimization means of claim 25 further comprises a means for creating VMs and assigning VMs to workloads in response to instructions received from the global resource allocator partitioning component.

28. A computer program product for use with a computer hosting architecture, for providing finer grain control in a virtual machine (VM) based hosting architecture, comprising:

a computer-readable medium

means, provided on the computer-readable medium, for identifying resource requirements for multiple different workloads in the VM based hosting architecture;

means, provided on the computer-readable medium, for assigning VMs from multiple server machines to a workload according to the identified resource requirements; and

means, provided on the computer-readable medium, for assigning resources at each of the multiple server machines to the assigned VMs according to the identified resource requirements.